International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

A Review on Effective Disease Classification Using Machine Learning Techniques on Healthcare Data

Rajat Mourya
Department of Computer Science and Engineering
Research Scholar, Rajshree Institute of Management and Technology, Bareilly
Dr Ruchin Jain (HOD)
Department of Computer Science and Engineering
Guide, Rajshree Institute of Management and Technology, Bareilly

Abstract
The integration of machine learning (ML) techniques into healthcare has emerged as a transformative approach for disease classification, offering improved diagnostic accuracy, speed, and efficiency. With the exponential growth of healthcare data from electronic health records, medical imaging, laboratory results, and wearable devices, traditional diagnostic methods face limitations in processing and interpreting large volumes of complex data. Machine learning models, particularly supervised learning algorithms such as Logistic Regression, Decision Trees, and Support Vector Machines, are increasingly being used to classify diseases including diabetes, cancer, heart disease, and neurological disorders. These models learn from historical data to identify hidden patterns and relationships between patient attributes and disease outcomes, supporting early detection and personalized treatment plans. Despite the promising capabilities, challenges remain, including data imbalance, lack of standardization, privacy concerns, and the need for model interpretability. Moreover, the successful deployment of ML models requires high-quality datasets and careful validation to ensure clinical relevance and reliability. This review explores the current landscape of ML-based disease classification, evaluates the performance of various algorithms, and highlights recent advancements in the field. It also discusses ongoing challenges and future directions for research. The study contributes to a better understanding of how ML can be leveraged to enhance healthcare outcomes through intelligent, data-driven disease prediction and diagnosis.
Keywords: Machine Learning, Disease Classification, Healthcare Data, Predictive Models, Diagnostic Accuracy

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

Introduction

With its rapid growing with the help of machine learning (ML) technologies, the field of healthcare has been drastically changed and at the core of it is disease classification and diagnosis. As the amount of healthcare data generated from electronic health records (EHRs), medical imaging, diagnostic laboratory results, and wearable devices increases exponentially, human diagnostic approaches are augmented by intelligent systems which are able to discern patterns and predict diseases with high accuracy. Supervised learning algorithms like various decision trees, support vector machines (SVM), k-nearest neighbor (KNN), deep learning networks are performing extremely well for the prediction and classification of such complex medical conditions such as cancer, diabetes, heart disease, neurological disorder, etc. Trained on labeled healthcare datasets, these models can pick up underlying relationship between symptoms, biomarkers and diagnosis better than traditional statistical methods. Besides, ML techniques integrated in processes also help clinicians with data-driven decision making, providing accurate results in a fast pace, optimizing patient outcomes and health care resource usage.

The objective of this review is to critically review some of the existing machine learning approaches applied in medical diagnosis of disease classification in healthcare. Strengths and weaknesses of different algorithms are analyzed, the nature of medical data, preprocessing methods, and evaluation metrics used in a given application for accuracy of the results are all also explored. It also emphasizes recent achievement and applications of ML in disease prediction, disease prediction, and personal treatment planning. Besides, key challenges, including data imbalance, interpretability of models, privacy concerns, and the requirement of standardized datasets, are discussed in the review. While healthcare has generally not yet come close to truly embracing digital transformation, the opportunity to understand and use the power of machine learning as a tool for robust, efficient and ethical diagnostic systems becomes more important for this domain. The goal of this review is to provide helpful insights about current trends and future directions for ML based disease classification, which will help researchers and practitioners from various areas exploit this dynamic and impactful field.

Need of the Study

With the rising complexity and magnitude of healthcare data, there is an imperative need for modern analytical tools that can be used to extract significant insights to achieve accurate and prompt disease diagnosis. This is because traditional diagnostic approaches are based on

heuristics enacted by experts, necessitating manual interpretation and static guidelines—all of which are prone to human error, inconsistency, and latency—particularly for resource-poor environments. With machine learning (ML), we can have a powerful alternative; we can leverage tools like this to analyze large dataset, patterns we have not identified, in real time to come up with real time predictions. Especially, this technological advancement is helpful in controlling the chronic disease, detecting the early signs of warning, and supporting the clinical decision-making process. The rise of gaining diagnostics accuracy, improved patient care and decreasing cost in health care using the data driven methods necessitates the need for this study. Furthermore, the greater availability of electronic health records, bio signal data, and wearable health devices boasts an ample bedrock to train and evolve ML models. While ML has been successfully applied to healthcare, it also has known challenges like lack of data quality, algorithm bias, lack of interpretability, and ethical concerns, which are worth investigating and understanding. This study is important for making an exploration of the effectiveness, reliability and applicability of different ML techniques in classifying diseases and to also discover the best practices, gaps that currently exist and finally directions to go in the future. This work is to support the researchers, healthcare professionals and policymakers in implementing ML based robust solutions for modern healthcare to be the more intelligent, more proactive, and more patient centered system through reviewing and analyzing current methodologies.

Applications of ML in Disease Classification

The incorporation of machine learning has facilitated the automation of a critical activity, which has contributed to significant improvement of accuracy and efficiency for disease classification from complex data in the healthcare setting. The most prominent application is in cancer diagnosis breast, lung and skin cancers. Some ML algorithms such as Support Vector Machine (SVM), Decision Trees, Convolutional Neural Networks (CNNs) have been able to classify tumors as either benign or malignant given biopsy result datasets, MRI datasets or datasets from gene expression profiles. CNNs, in particular, have performanced well with radiology image analysis such as mammograms and CT scans where the results are fast and consistent compared to manual assessment.

There is another major area of application of prediction to chronic diseases such as diabetes and cardiovascular disorders. Clinical parameters like blood pressure, glucose levels, cholesterol and family history are used to predict the onset or progression of these diseases

using algorithms such as Random Forests, K Nearest Neighbors (KNN) and Logistic Regression. Early identification of the at risk individual can be facilitated with the help of these models, which allows timely intervention and prevents complication. Wearable devices and mobile health apps, in turn, are instrumented with predictive models to continuously monitor and alert regarding abnormal patterns.

Machine learning has been applied to and been successful in the field of neurological and mental health for classifying conditions such as Alzheimer's, Parkinson's disease, and depression. The speech patterns and behavioral data and patient history of the patient is obtained and analyzed using natural language processing (NLP) techniques along with ML models to detect the early signs of cognitive decline. Similar EEG and MRI data are also processed using deep learning models for discriminating healthy from affected individuals. Early diagnosis of this disease is made possible, and individualized treatment planning is supported that leads to significantly improved patient quality of life.

Infectious disease detection is also another area where ML plays a very important role such as for example COVID 19, tuberculosis and influenza. Models can classify disease stage by comparing patient symptoms and lab test results, as well as predict how likely an infection is. During the COVID 19 pandemic, Machine Learning Models helped in triaging patients, predicting ICU Admissions, and estimating the gravity of the situation via chest X rays and by analyzing the clinical indicators. The real-time applications carried out highlight the need of machine learning to support fast, scalable and accurate disease classification in the routine and emergency health environments.

Different Types of Machine Learning Techniques

Disease classification is a task where machine learning techniques help to a great extent because it allows the system to learn from healthcare data and predict with great accuracy. However, among these, supervised learning models are the most frequently used as they need to be trained via labeled dataset to train algorithms used for classifying diseases. They are able to identify the patterns and relationships in the complex data sets so it is very good to use these models in medical diagnostics. In this section we will focus on 3 widely used Machine Learning techniques in healthcare namely Logistic Regression, Decision Tree Classifier and Support Vector Machine Classifier (SVM). They have their own strengths and different applications depending on the nature of the data and the classification problem to be solved.

Logistic Regression

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

Logistic Regression is a type of statistical model that is used for binary classification problem (the output has two possible outcomes) like diseased or not diseased. It uses the logistic (sigmoid) function to calculate a probability of a particular outcome using the input features. Logistic Regression is employed much in healthcare to predict the disease (diabetes, heart disease or cancer) from the presence or absence of the disease and the clinical parameters (age, blood pressure, glucose level and cholesterol). The model is simple, easy to interpret, has low computational complexity, and is deemed a good first option to assess risk of disease. While being adaptable is not challenging, the RW may not perform well when the curve created between input features and output is nonlinear or complex.

Decision Tree Classifier

The decision Tree is a flow chart like model that by making a series of questions about the input features decides what input features to use. It divides data in the subsets according to feature value and then gives the final classification. Decision Trees are highly intuitive and support visual representation due to which they are easily interpreted by healthcare professionals. Symptoms and medical history is useful to diagnoses multiple diseases depending on the symptom or the medical history. While overly prone to overfitting on noisy data, Random Forest ensemble methods can be used to address this problem.

Support Vector Machine

The Support Vector Machine (SVM) is a strong classification algorithm that attempts to find a hyperplane that best separates data groupings into different classes using the maximum margin. It has been particularly effective in high dimensional spaces, and was very widely used in the context of medical imaging and genomics. Kernel functions allow SVM to cope with both the linear and non-linear data. This method is appropriate for classifying a disease like cancer and neurological disorder, because of its high accuracy and handling complex datasets. SVMs, however, are highly computationally intensive and depend on tuning some parameters carefully.

Literature Review

Karna Vishnu Vardhana Reddy et al (2021) Despite being among those leading causes of death worldwide, early detection is very important for an effective treatment and prevention of heart disease. In this research, PCA was used to reduce the dataset dimensionality, remove the redundant and the less important features for better model performance and complexity reduction. I apply various ML algorithms (i.e. Logistic Regression, Random Forest, Support

Vector Machine) to the processed data to identify the patterns and factors that are responsible for heart disease. The study combines PCA with these predictive models to have better generalization and classification accuracy. It shows that dimensionality reduction using PCA leads to more reliable and interpretable prediction that enables medical professionals make better diagnostic decision.

M.Ganesan and Dr. N. Sivakumar (2019) Wearable sensors are IoT devices that continuously collect the vital physiological data, including heart rate, blood pressure, ECG signals and oxygen levels. However, this real-time data is transported to a cloud-based platform where the data is analyzed using Machine Learning algorithms to predict the risk of having a heart disease. Techniques including Decision Trees, Support Vector Machines, and Neural Networks are used to break abnormal patterns and stage alerts for medical treatment. Fundamentally, IoT integration with machine learning makes it possible to monitor health continuously, diagnose timely and provide personalized care. This system is not only better for diagnostic accuracy but also enables remote healthcare delivery, lowering hospital visits, and eventually, is better at proactive and efficient patient management.

Priyan Malarvizhi Kumar (2018) In this article, we have proposed the architecture that runs on wearable IoT sensors for continuous monitoring of heart rate, ECG, blood pressure, and oxygen saturation. The secure data are sent to a cloud based server, where machine learning algorithms work on the inputs in real time to catch the potential cardiac abnormalities. It is intended to do with high accuracy and low latency so that it could timely alert and provide medical responses. We utilize machine learning models like Random Forest, Gradient Boosting and Deep Neural Networks due to their strong ability for classification and prediction. They can share data through the intelligent IoT architecture to remotely monitor patients, reduce the dependence on frequent clinical visits and remotely provide proactive healthcare by tracking and identifying the risk at an earlier stage supporting better patient outcome and reduction in mortality.

Prabal Verma, Sandeep K. Sood (2018) An advanced healthcare model based on cloud computing and Internet of Things (IoT) technology, where a disease diagnosis is performed efficiently and accurately. Within such a framework, IoT enabled wearable sensors and smart medical equipment in the network collect real time physiological data from the patients such as heart rate, body temperature, blood pressure, oxygen levels etc. The data is transmitted to a centralized cloud platform that stores, processes and analyzes the data with the help of

intelligent algorithms and machine learning models. Using the cloud centric approach, the data can be accessible at high speeds, managed by healthcare providers to scale quickly, integrated well with electronic health records, and monitor patients remotely and make decisions with data in a more timely way. Early detection of several diseases in a system that also improves patient engagement while reducing healthcare infrastructure burden, the proposed framework promotes improvement in means of healthcare delivery by providing continuous monitoring, secure data management, and personalized diagnosis at low cost and accessibility.

Ashwini Shetty, Naik, C. (2016. Heart conditions are complex and of multi factorial nature making data mining a promising tool to extract valuable information from large health care datasets. In the research, a number of popular approaches such as Naive Bayes, Decision Trees, Support Vector Machines, K-Nearest Neighbors and Neural Networks were evaluated for their ability to classify and predict the occurrence of heart disease utilizing clinical parameters such as age, cholesterol level, blood pressure and ECG results. These help uncover hidden correlation and trends that might not be noticeable through the traditional way of analysis. Results of various comparisons revealed the strengths and shortcomings of each of the models by accuracy, precision, recall and computational efficiency. The study applies these data mining methods and proves to have benefited diagnostic support in preventing and treating heart disease faster and more wisely with healthcare providers.

Aydin, S. (2016) According to the study By heart disease being the leading cause of death worldwide, early and accurate detection is invaluable. In this research, multiple data mining techniques like Decision Trees, Naive Bayes, K-Nearest Neighbors, Support Vector Machines and Artificial Neural Networks are applied and compared to the clinical datasets that have features like age, gender, cholesterol levels, blood pressure and ECG results. Each technique is evaluated by key metrics: accuracy, sensitivity, specificity, precision and F1score. The comparative analysis finds that the strengths and weaknesses of the two methods, as well as the models suitable in the prediction of the occurrence of heart disease, relative to different contexts. The study identifies the most reliable and efficient algorithms and offers useful information for the creation of such tools as an aid for healthcare professionals to make timely and data-driven decisions regarding medical treatment.

Bayasi, N. and Tekeste (2016) Specifically, the processor is designed for use in the portable and possibly wearable health monitoring devices where low power consumption is crucial. It combines advanced signal processing methods and machine learning algorithms to

automatically detect abnormal heart rhythms in real time by means of QRS complex, heart rate variability, and morphological patterns. The system allows for continuous monitoring without the need of continuous connectivity to the cloud by embedding the prediction capability directly inside the processor, which shortens the response time and protects the privacy of the patient. A design for a multi–core imaging inertial broadside system designed to guarantee high accuracy and low latency while occupying a small physical area and remains energy efficient is proposed. By supporting proactive cardiac care, this innovation permits timely intervention by doctors and curbs the chances of sudden cardiac events among the at risk individuals.

Berikol, B. and Yildiz (2016) This paper describes the development of Support Vector Machine (SVM) algorithms for accurately diagnosing Acute Coronary Syndrome (ACS), which is a severe and potentially deadly cardiac disease. ACS covers a population of urgent heart diseases such as unstable angina and myocardial infarction, for which angina, rapid and precise diagnosis is needed. The SVM classifier is trained with the clinical and physiological data such as chest pain characteristics, ECG readings, troponin levels, heart rate and blood pressure. Since the SVM becomes strong in a nonlinear observation and a high dimensional space, it allows fitting complex medical datasets. Supervised learning allows the model to identify such patterns and boundaries that separate ACS from other cardiac and non-cardiac conditions. The results show that SVMs are a highly accurate, sensitive and specific diagnostic method for the early detection. This helps clinicians in making timely decisions and improving patient outcomes and reduces treatment delays.

Chebbi, A. (2016) Early prediction of heart disease is of paramount importance in reducing mortality and improving patient care, because heart disease is today a major global health problem. For this research, I use the data of parameters like age, gender, blood pressure, cholesterol levels, chest pain type, and ECG results to train and test the predictive models. Through application of Decision Trees, Naive Bayes, K – Nearest Neighbors, and Support Vector Machines techniques, data mining techniques are used to mine those data to discover not apparent patterns and correlations, which may reveal, presence or risk of heart disease. These methods are evaluated in terms of accuracy, sensitivity and precision. Results indicate that data mining has great potential for enhancing diagnostic outcomes and providing such solutions can assist the healthcare professionals in making well informed choices. In the end, the proposed research aims at using intelligent and data driven solutions in modern medical diagnosis systems.

International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)
An International Open Access, Peer-Reviewed Refereed Journal
Impact Factor: 6.4 Website: https://ijarmt.com ISSN No.: 3048-9458

Research problem

The large number of healthcare data, due to the widespread use of electronic records, wearable devices, and diagnostic technologies, offers both an opportunity and challenge to disease diagnosis and classification. However, such data present a challenge to the prevalent diagnostic methods and thus result in delays, inaccuracies, and inconsistencies within clinical decision making. This section takes into account the research problem that how well the machine learning (ML) techniques can be used to classify diseases based on different healthcare datasets? Despite many ML algorithms being introduced and used for different medical conditions, there is no systematic comparison of performance, suitability and limitation of these approaches in real world clinical scenarios. However, these systems face challenges like data imbalance, lack of interpretability, privacy issues and unsuccessful integration into existing healthcare systems, thereby limiting their practical deployment. In this review, we discuss the strengths and weaknesses of techniques that have been often used for disease classification, namely: Logistic Regression, Decision Trees, Support Vector Machines and Deep Learning. This paper proposes to guide future research and implementation efforts towards development of robust, accurate, and interpretable ML-based diagnostic tools that can improve the quality of healthcare delivery and patient outcomes through discussing recent advances and existing gaps.

Conclusion

Machine learning techniques have recently been integrated into healthcare, and became able to improve disease classification with the help of fast, accurate and scalable solutions. Machine learning models can find patterns that humans often find difficult to see by using various data sources, including electronic health records, medical imaging, and biosignals. In addition to improving diagnostic accuracy, these models help early detection and make prediction of the risk, which are important for providing effective treatment and better patient outcomes. Some of these diseases include diabetes and cardiovascular diseases, but also complex diseases that are difficult to treat, such as cancer and others like neurological disorders.

Despite its promising applications, the implementation of machine learning in healthcare still faces several challenges. Others include privacy, transparency of algorithms and ability to interpret results, as well as the desire for a standardized dataset. Moreover, there are more workflows that ML models need to be properly and safely integrated into, and appropriately validated and subject to regulatory oversight. However, limitations in deep learning,

explainable AI, and federated learning are being addressed through ongoing progress in these fields, leading to more robust and trustworthy systems. This review documents advances in the use of ML for disease classification and stresses the need for further interdisciplinary research to harvest future ethical, effective, and patient-engaged healthcare technologies.

References

[1] Karna Vishnu Vardhana Reddy, Irraivan Elamvazuthi, Azrina Abd Aziz, Sivajothi Paramasivam and Hui Na Chua, “Heart Disease Risk Prediction using Machine Learning with Principal Component Analysis”, International Conference on Intelligent and Advanced Systems (ICIAS), pp. 01-05, IEEE 2021.

[2] M. Ganesan and Dr. N. Sivakumar, “IoT based heart disease prediction and diagnosis model for healthcare using machine learning models”, International Conference on System, Computation, Automation and Networking (ICSCAN), pp. 01-05, IEEE 2019.

[3] Priyan Malarvizhi Kumar, Usha Devi Gandhi, “A novel Internet of Things architecture with machine learning algorithm for early detection of heart diseases", Computers and Electrical Engineering, Vol.65, pp. 222–235, 2018.

[4] Prabal Verma, Sandeep K. Sood, "Cloud-centric IoT based disease diagnosis healthcare framework", Journal of Parallel Distribution Computer, Vol. 116, Issue 06, pp. 27-38, 2018.

[5] Amin Khatami and Abbas Khosravi, “Medical image analysis using wavelet transform and deep belief networks”, Journal of Expert Systems with Applications, Vol. 3, Issue 4, pp. 190–198, 2017.

[6] Sanjay Kumar Sen, "Predicting and Diagnosing of Heart Disease using Machine Learning Algorithms", International Journal of Engineering and Computer Science (IJECS), Vol. 6, Issue 7, pp. 21623-21631, 2017.

[7] X Liu, X Wang, Q Su M Zhang and Y Zhu Q Wang, “A hybrid classification system for heart disease diagnosis based on the RFRS method", Computational and mathematical methods in medicine, pp. 01-11, 2017.

[8] Ashwini Shetty, Naik, C., “Different data mining approaches for predicting heart disease”, International journal of innovative research in science, engineering and technology, Vol. 3, Issue 2, pp. 277–281, 2016.

[9] Aydin, S., "Comparison and evaluation data mining techniques in the diagnosis of heart disease", Indian journal of science and technology, Vol. 6, Issue 1, pp. 420–423, 2016.

[10] Bayasi, N. and Tekeste, "Low-power ECG-based processor for predicting ventricular arrhythmia", Journal of IEEE transactions on very large scale integration systems, Vol. 24, Issue 5, pp. 1962–1974, 2016.

[11] Berikol, B. and Yildiz, "Diagnosis of acute coronary syndrome with a support vector machine", Journal of Medical System, Vol. 40, Issue 4, pp. 11–18, 2016.

[12] Chebbi, A., 'Heart disease prediction using data mining techniques', International journal of research in advent technology, Vol. 25, Issue 3, pp. 781–794, 2016.